Audio-Video Integration for Background Modelling

Marco Cristani, Manuele Bicego, and Vittorio Murino

Dipartimento di Informatica, University of Verona
Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy
{cristanm,bicego,murino}@sci.univr.it

Abstract. This paper introduces a new concept of surveillance, namely, audio-visual data integration for background modelling. Actually, visual data acquired by a fixed camera can be easily supported by audio information allowing a more complete analysis of the monitored scene. The key idea is to build a multimodal model of the scene background, able to promptly detect single auditory or visual events, as well as simultaneous audio and visual foreground situations. In this way, it is also possible to tackle some open problems (e.g., the sleeping foreground problems) of standard visual surveillance systems, if they are also characterized by an audio foreground. The method is based on the probabilistic modelling of the audio and video data streams using separate sets of adaptive Gaussian mixture models, and on their integration using a coupled audio-video adaptive model working on the frame histogram, and the audio frequency spectrum. This framework has shown to be able to evaluate the time causality between visual and audio foreground entities. To the best of our knowledge, this is the first attempt to the multimodal modelling of scenes working on-line and using one static camera and only one microphone. Preliminary results show the effectiveness of the approach at facing problems still unsolved by only visual monitoring approaches.

1 Introduction

Automated surveillance systems have acquired an increased importance in the last years, due to their utility in the protection of critical infrastructures and civil areas. This trend has amplified the interest of the scientific community in the field of the video sequence analysis and, more generally, in the pattern recognition area [1]. In this context, the most important low-level analysis is the so called background modelling [2,3], aimed at discriminating the static scene, namely, the background (BG), from the objects that are acting in the scene, i.e., the foreground (FG). Despite the large related literature, there are many problems that are still open [3], like, for instance, the sleeping foreground problem. In general, almost all of the methods work only at the visual level, hence resulting in *video* BG modelling schemes. This could be a severe limitation, since other information modalities are easily available (e.g., audio), which could be effectively used as complementary information to discover "activity patterns" in a scene.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3022, pp. 202–213, 2004.

In this paper, the concept of *multimodal*, specifically audio-video, BG modelling is introduced, which aims at integrating different kinds of sensorial information in order to realize a more complete BG model. In the literature, the integration of audio and visual cues received a growing attention in the last few years. In general, audio-visual information have been used in the context of speech recognition, and, recently, of scene analysis, especially person tracking. A critical review of the literature devoted to audio-video scene analysis is reported in section 2.

In order to integrate audio and visual information, different adaptive BG mixture models are first designed for monitoring the segregated sensorial data streams. The model for visual data operates at two levels. The first is a typical time-adaptive per-pixel mixture of Gaussians model [2], able to identify the FG present in a scene. The second model works on the FG histogram, and is able to classify different FG events. Concerning the audio processing scheme, the concept of *audio* BG modelling is introduced, proposing a system able to detect unexpected audio events. In short, a multiband frequency analysis was first carried out to characterize the monaural audio signal, by extracting features from a parametric estimation of the power spectral density. The audio BG model is then obtained by modelling these features using a set of adaptive mixtures of Gaussians, one for each frequency subband.

Concerning the on-line fusion of audio information with visual data, the most basic issue to be addressed is the concept of "synchrony", which derives from psycho-physiological research [4,5]. In this work, we consider that visual and audio FG that appear "simultaneously" are synchronous, i.e., likely causally correlated. The correlation augments if both FG events persist along time.

Therefore, a third module based on adaptive mixture models operating on audio-visual data has been devised. This module operates in a hybrid space composed by the audio frequency bands, and the FG histogram bins, allowing the binding of concomitant visual and audio events, which can be labelled as belonging to the same multimodal FG event. In this way, a globally consistent multilevel probabilistic framework is developed, in which the segregated adaptive modules control the different sensorial audio and video streams separately, and the coupled audio-video module monitors the multimodal scenario to detect concurrent events. The three modules are interacting each other to allow a more robust and reliable FG detection.

In practice, our structure of BG modelling is able to face serious issues of standard BG modelling schemes, e.g., the sleeping FG problem [3].

The general idea is that an audio-visual pattern can remain an actual FG even if one of the components (audio or video) is missing. The crucial step is therefore the discovery of the audio-visual pattern in the scene.

In summary, the paper introduces several concepts related to the multimodal scene analysis, discussing the involved problems, showing potentialities and possible future directions of the research. The key contributions of this work are: 1) the definition of the novel concept of multimodal background model, introducing, together with video data, audio information processing performing an

auditory scene analysis using only *one* microphone; 2) a method for integrating audio and video information in order to discover synchronous *audio-visual* patterns *on-line*; 3) the implementation of these audio-visual fusion principles in a probabilistic framework working on-line and able to deal with complex issues in video-surveillance, i.e., the sleeping foreground problem.

The rest of the paper is organized as follows. In Section 2, the state of the art related to the audio-video fusion for scene analysis is presented. The whole strategy is proposed in Section 3, and preliminary experimental results are reported in Section 4. Finally, in Section 5, conclusions are drawn.

2 State of the Art of the Audio-Visual Analysis

In the context of audio-visual data fusion it is possible to individuate two principal research fields: the on-line audio-visual association for tracking tasks, and the more generic off-line audio-visual association, in which the concept of audio-visual synchrony is particularly stressed.

In the former, the typical scenario is an indoor known environment with moving or static objects that produce sounds, monitored with fixed cameras and fixed acoustic sensors. If an entity emits sound, the system provides a robust multimodal estimate of the location of the object by utilizing the time delay of the audio signal between the microphones and the spatial trajectory of the visual pattern [6,7]. In [6], the scene is a conference room equipped with 32 omnidirectional microphones and two stereo cameras, in which a multi-object 3D tracking is performed. With the same environmental configuration, in [8] an audio source separation application is proposed: two people speak simultaneously while one of them moves through the room. Here the visual information strongly simplifies the audio source separation.

The latter class of approaches employs only one microphone. In this case the explicit notion of the spatial relationship among sound sources is no more recoverable, so the audio-visual localization process must depend purely on the concept of synchrony, as stated in [9]. Early studies about audio-visual synchrony comes from the cognitive science. Simultaneity is one of the most powerful cues available for determining whether two events define a single or multiple objects; moreover, psychophysical studies have shown that the human attention focuses preferably on sensory information perceived coupled in time, suppressing the others [4]. Particular effort is spent in the study of the situation in which the inputs arrive through two different sensory modalities (such as sight and sound) [5].

Most of the techniques in this context make use of measures based on the mutual information criterion [8,10]. These methods extract the pixels of the video sequences that are most related to the occurring audio data using maximization of the mutual information between the entire audio and visual signals, resulting therefore in an off-line processing. For instance, they are used for video-conference annotation [10]: audio and video features are modelled as Gaussians processes, without a distinction between FG and BG. The association is exploi-

ted by searching for a correlation in time of each pixel with each audio feature. The main problem is that it assumes that the visual pattern remains fixed in space; further, the analysis is carried out completely off-line.

The method proposed in this paper tries to bridge these two research areas. To the best of our knowledge, the proposed system constitutes the first attempt to design an on-line *integrated audio-visual* BG modelling scheme using only one microphone, and working in a loosely constrained environment.

3 The Audio-Video Background Modelling

The key methodology is represented by the on-line time-adaptive mixture of Gaussians method. This technique has been used in the past to detect changes in the grey level of the pixels for background modelling purposes [2]. In our case, we would like to exploit this method to detect *audio foreground*, video foreground objects, and *joint audio-video* FG events, by building a robust and reliable *multimodal* background model. The basic concepts of this approach are summarized in Section 3.1. The customization in the case of visual and audio background modelling is presented in Section 3.2, and in Section 3.3, respectively. Finally, the integration between audio and video data is detailed in Section 3.4, and how the complete system is used to solve a typical problem of visual surveillance system is reported in Section 3.5.

3.1 The Time-Adaptive Mixture of Gaussians Method

The Time-Adaptive mixture of Gaussians method aims at discovering the deviance of a signal from the expected behavior in an on-line fashion. A typical video application is the well-know BG modelling scheme proposed in [2]

The general method models a temporal signal with a time-adaptive mixture of Gaussians. The probability to observe the value $z^{(t)}$, at time t, is given by:

$$P(z^{(t)}) = \sum_{r=1}^{R} w_r^{(t)} \mathcal{N}\left(z^{(t)} | \mu_r^{(t)}, \sigma_r^{(t)}\right) \tag{1}$$

where $w_r^{(t)}$, $\mu_r^{(t)}$ and $\sigma_r^{(t)}$ are the mixing coefficients, the mean, and the standard deviation, respectively, of the r-th Gaussian of the mixture associated to the signal at time t. At each time instant, the Gaussians in a mixture are ranked in descending order using the w/σ value. The R Gaussians are evaluated as possible match against the occurring new signal value, in which a successful match is defined as a pixel value falling within 2.5σ of one of the component. If no match occurs, a new Gaussian with mean equal to the current value, high variance, and low mixing coefficient replaces the least probable component.

If r_{hit} is the matched Gaussian component, the value $z^{(t)}$ is labelled as unexpected (i.e., foreground) if $\sum_{r=1}^{r_{hit}} w_r^{(t)} > T$, where T is a threshold representing

the minimum portion of the data that supports the "expected behavior". The evolution of the components of the mixtures is driven by the following equations:

$$w_r^{(t)} = (1-\alpha)w_r^{(t-1)} + \alpha M^{(t)}, 1 \leq r \leq R, \tag{2}$$

where $M^{(t)}$ is 1 for the matched Gaussian (indexed by r_{hit}), and 0 for the others; α is the adaptive rate that remains fixed along time. It is worth to notice that the higher the adaptive rate, the faster the model is "adapted" to scene changes.

The μ and σ parameters for unmatched Gaussians remain unchanged, but, for the matched Gaussian component r_{hit}, we have:

$$\mu_{r_{hit}}^{(t)} = (1-\rho)\mu_{r_{hit}}^{(t-1)} + \rho z^{(t)} \tag{3}$$

$$\sigma_{r_{hit}}^{2\,(t)} = (1-\rho)\sigma_{r_{hit}}^{2\,(t-1)} + \rho \left(z^{(t)} - \mu_{r_{hit}}^{(t)}\right)^T \left(z^{(t)} - \mu_{r_{hit}}^{(t)}\right) \tag{4}$$

where $\rho = \alpha \mathcal{N}\left(z^{(t)} | \mu_{r_{hit}}^{(t)}, \sigma_{r_{hit}}^{(t)}\right)$.

3.2 Visual Foreground Detection

One of the goal of this work is to detect untypical video activity patterns starting simultaneously with audio ones. In order to discover these visual patterns, a video processing method has been designed, which is composed by two modules: a standard per-pixel FG detection module, and an histogram-based novelty detection module.

The former is realized using the model introduced in the previous section in a standard way [2], in which the processed signal $z^{(t)}$ is the time evolution of the gray level. We use a set of independent adaptive mixtures of Gaussians, one for each pixel. In this case, an unexpected valued pixel $z_{uv}^{(t)}$ (where u, v are the coordinates of the image pixel) is the visual *foreground*, *i.e.*, $z_{uv}^{(t)} \in FG$. Please, note that all mixtures' parameters are updated with a fixed learning coefficient $\tilde{\alpha}$.

The latter module is also realized using the time-adaptive mixture of Gaussians method, using the same learning rate $\tilde{\alpha}$ of the former module, but in this case we focus on the histogram of those pixels classified as foreground. The idea is to compute at each step the gray level histogram of the FG pixels and associating an adaptive mixture of Gaussian to each bin, looking for variations of the bin's value. This means that we are monitoring the number of pixels of the foreground that have a specific gray value. If the number of pixels associated to the foreground *grows*, i.e., some histogram bins increase their values, then an object is appearing in the scene, otherwise is disappearing. We choose to monitor the histogram instead of the number of FG pixels directly (which can be in principle sufficient to detect new objects), as it allows the discrimination between different FG objects, and in order to detect audio-visual patterns composed by single objects. We are aware that this simple characterization leaves some ambiguities (e.g., two equally colored objects are not distinguishable, even if the impact of this problem may be weakened by increasing the number of bins), but

this representation has the appealing characteristic of being invariant to spatial localization of the foreground, which is not constrained to be statically linked to a spatial location (as in other audio-video analysis approaches)[1].

3.3 Audio Background Modelling

The audio BG modelling module aims at extracting information from audio patterns acquired by a *single* microphone. In the literature, several approaches to audio analysis are present, mainly focused on the computational translation of psychoacoustics results. One class of approaches is the so called "computational auditory scene analysis"(CASA) [12], aimed at the separation and classification of sounds present in a specific environment. Closely related to this field, but not so investigated, there is the "computational auditory scene recognition" (CASR) [13,14], aimed at environment interpretation instead of analyzing the different sound sources.

Besides various psycho-acoustically oriented approaches derived from these two classes, a third approach tried to fuse "blind" statistical knowledge with biologically driven representations of the two previous fields, performing audio classification and segmentation tasks [15], and source separation [16,17] (blind source separation). In this last approach, many efforts are devoted in the speech processing area, in which the goal is to separate the different voices composing the audio pattern using several microphones [17] or only one monaural sensor [16].

The approach presented in this paper could be inserted in this last category: roughly speaking, we implement a multiband spectral analysis on the audio signal at video frame rate, extracting energy features from $a_1, a_2, \ldots, a_M$ frequency subbands. More in detail, we subdivide the audio signal in overlapped temporal window of fixed length W_a, in which each temporal window ends at the instant corresponding to the t-th video frame, as depicted in Fig.1. For each window, a parametric estimation of the power spectral density with the Yule-Walker Auto Regressive method [18] is performed. In this way, an estimation $a_i^{(t)}$ of the spectral energy relative to the interval $[t-W_a, t]$ is obtained for the i-th subband, $i = 1, 2, \ldots, M$. These features have been chosen as they are able to discriminate between different sound events [13]; further, they can be easily computed at an elevate temporal rate.

As typically considered [16], the energy during time in different frequency bands can transport independent information. Therefore, we instantiate one time-adaptive mixture of Gaussians for each band of the frequency spectrum. Also in this case, all mixtures' parameters are updated with a fixed learning coefficient $\tilde{\alpha}$, equal to the one used for the video channel. In this way, we are able to discover unexpected audio behaviors for each band, indicating an audio foreground.

[1] Actually, more sophisticated tracking approaches based on histograms have already been proposed in literature [11], and are subjects of future work.

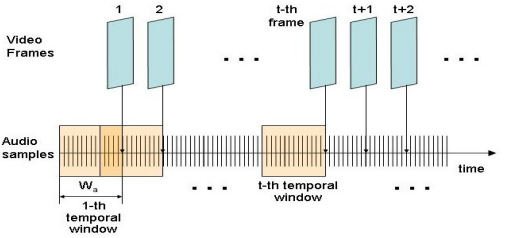

Fig. 1. Organization of the multimodal data set: at each video frame, an audio signal analysis is carried out using a temporal window of length W_a

3.4 The Audio-Visual Fusion

The audio and visual spaces are now partitioned in different independent subspaces, the audio subbands $a_1, a_2, \ldots, a_M$, and the video FG histogram bins $h_1, h_2, \ldots, h_N$, respectively, in which independent FG monomodal patterns may occur. Therefore, given an audio subband $a_i^{(t)}$, and a video histogram bin $h_j^{(t)}$ at time t, we can define an history of the mono-modal FG patterns $A_i^{(t)}$, $i = 1, \ldots, M$, and $H_j^{(t)}$, $j = 1, \ldots, N$, as the patterns in which the values of a given component of the $i-th$ mixture for the audio, and the $j-th$ mixture for the video are detected as foreground along time. Formally, let us denote $A_i^{(t)}$ and $H_j^{(t)}$ as:

$$A_i^{(t)} = [a_i^{(t_{q,i})}, a_i^{(t_{q,i}+1)}, \ldots, a_i^{(t)} \in FG] \tag{5}$$

$$H_j^{(t)} = [h_j^{(t_{u,j})}, h_j^{(t_{u,j}+1)}, \ldots, h_j^{(t)} \in FG] \tag{6}$$

where $t_{q,i}$ is the first instant at which the $q-th$ Gaussian component of the audio mixture of the $i-th$ sub-band becomes FG, and the same applies for $t_{u,j}$ related to the video data. Clearly, $A_i^{(t)}$ and $H_j^{(t)}$ are possibly not completely overlapped, so $t_{q,i}$ in general can be different from $t_{u,j}$. Therefore, in order to evaluate the degree of concurrency, we define a *concurrency* value as $\beta_{i,j} = |t_{q,i} - t_{u,j}|$. Obviously, the higher this value, the weaker the synchronization.

As previously stated, the synchronism gives a natural *causal* relationship for processes coming from different modalities [4]. In order to evaluate this causal dependency along time, we state as highly correlated those concurrent audio-video FG patterns explaining, in their jointly evolution, a nearly stable behavior. Consequently, we couple all the audio FG values with all the visual FG values occurring at time step t, building an $M \times N$ *audio-visual FG matrix* $AV^{(t)}$, where

$$AV^{(t)}(i,j) = \begin{cases} (a_i^{(t)}, h_j^{(t)}) & \text{if } \; a_i^{(t)} \in FG \bigwedge h_j^{(t)} \in FG \\ empty & \text{otherwise} \end{cases} \tag{7}$$

This matrix gives a snapshot of the degree of synchrony between audio and visual FG values, for all i, j. If $AV^{(t)}(i,j)$ is not empty, probably, $A_i^{(t)}$ and $H_i^{(t)}$ are in

some way synchronized. In this last case, we choose to model the evolution of these values using an on-line 2D adaptive Gaussian model. Therefore, at each time step t, we can evaluate the probability to observe a pair of audio-visual FG events, $AV^{(t)}(i,j)$, as

$$P(AV^{(t)}(i,j)) = \sum_{r=1}^{R} w_{AV_r}^{(t,i,j)} \mathcal{N}\left(AV^{(t)}(i,j) | \mu_r^{(t,i,j)}, \mathbf{\Sigma}_r^{(t,i,j)}\right) \tag{8}$$

Intuitively, the higher the value of the weight $w_{AV_r}^{(t,i,j)}$ matched by the observation $(a_i^{(t)}, h_j^{(t)})$ at time t, namely $w_{AV_{r_{hit}}}^{(t,i,j)}$, the more stable are the coupled audio-visual FG values along time, and it is more probable that a causal relation is present between audio and visual FG.

All the necessary information to assess the synchrony and the stability of a pair of audio and video FG patterns is now available. Therefore, a modulation of the evolution process of the 2D Gaussian mixture model is introduced in order to give more importance to a match with a couple of FG values belonging to likely synchronized audio and video patterns. We would like to impose that the higher the concurrency, the faster the stability of an AV value must be highlighted. In formulas, omitting the indices i, j for clarity

$$w_{AV_r}^{(t)} = (1 - \alpha_{AV}) w_{AV_r}^{(t-1)} + \alpha_{AV} M_{AV}^{(t)}, \quad 1 \leq r \leq R, \tag{9}$$

where

$$M_{AV}^{(t)} = \begin{cases} \frac{1}{\beta_{i,j}+1} = \frac{1}{|t_q - t_u|+1} & \text{for the matched 2D Gaussian} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

This equation [2] implies that if the synchronization does not occur at the same instant, the weight grows more slowly, and viceversa.

In order to subsume the concurrency and the stability behavior of the multimodal FG patterns, we finally introduce the *causality matrix* $\Gamma^{(t)} = [\gamma_{i,j}^t]$, for all $i = 1, \ldots, M$, and $j = 1, \ldots, N$, where

$$\gamma^{(t)}(i,j) = w_{AV_{r_{hit}}}^{(t,i,j)} \tag{11}$$

where $w_{AV_{r_{hit}}}^{(t,i,j)}$ is the weight of the 2D Gaussian component of the model matched by the pair of FG values $(a_i^{(t)}, h_j^{(t)})$.

As we will see in the experimental session, this model well describe the stability degree of the audio-visual FG, in an on-line unsupervised fashion.

3.5 Application to the Sleeping Foreground Problem

The sleeping foreground problem occurs when a moving object, initially detected as foreground, stops, and becomes integrated in the background model after a

[2] Any function inversely proportional to $\beta_{i,j}$ could be used; actually, different function choices do not sensibly affect the method performances.

certain period. We want to face this situation, under the hypothesis that there is a multimodal FG pattern, i.e. detecting the correlation between audio and video FG. In this situation, we maintain as foreground both the visual appearance of the object and the audio pattern detected, as long as they are present and stable in time. Technically speaking, we compute the learning rate of the mixture of Gaussians associated to the video histogram's bin j

$$\alpha_j^{(t)} = \min\left(\tilde{\alpha}, 1 - \max_i \gamma^{(t)}(i,j)\right) \tag{12}$$

where $\tilde{\alpha}$ is the learning rate adopted for both the segregated sensorial channels. The learning rates of the adaptive mixtures of all pixels which gray level belongs to the histogram bin j become $\alpha_j^{(t)}$. Moreover, also the learning rate of the mixture associated to the band $\arg\max_i \gamma^{(t)}(i,j)$ becomes $\alpha_j^{(t)}$. This measure implies that the most correlated audio FG pattern with the $j-th$ video FG pattern guides the evolution step, and viceversa. In practice we can distinguish $\min(M,N)$ different audio-video patterns. This may appear a weakness of this method, but this problem may be easily solved by using a finer discretization of the audio spectral, and of the histogram spaces. Moreover, other features could be used for the video data modelling, like, for instance, color characteristics.

4 Experimental Results

An indoor audio-visual sequence is considered, in which two sleeping FG situations occur: the former is associated with audio cues, and the latter is not. We will show that our system is able to deal with both situations.

More in detail, the sequence is captured at 30 frames per second, and the audio signal is sampled at 22.050 Hz. The temporal window used for multiband frequency analysis is equal to 1 second, and the order of the autoregressive model is 40. We undersample the 128×120 video image in a grid of 32×30 locations. Finally, we use 12 bins for the FG color histogram. Analogously, we perform spectral analysis using $M = 16$ logarithmic spaced frequency subbands, in which the frequency is measured in radians in the range $[0, \pi]$, and the power is measured in Decibel. As a consequence, we have an audio-visual space quantized in $M \times N = 16 \times 12$ elements. All adaptive mixtures are composed by 4 Gaussian components, and the learning parameter for the AV mixtures is fixed to 0.05, and for the separated channels $\tilde{\alpha}$=0.005, initially.

We compare our results with those proposed by an "only video" BG modelling, choosing as reference the standard video BG modelling adopted in [2], showing: 1) the resulting analysis of both BG modelling schemes; 2) the audio BG modelling analysis; 3) the histogram FG modelling analysis, able to individuate the appearance of new visual FG in the scene, and 4) the causality matrix, ordered by audio subbands per video histogram bins, that explains intuitively the intensity causal relationship in the joint audio-visual space.

As one can observe in Fig.2, at frame 50 both per-pixel BG modelling schemes locates a FG entering in the scene. This causes a strong increment in the gray

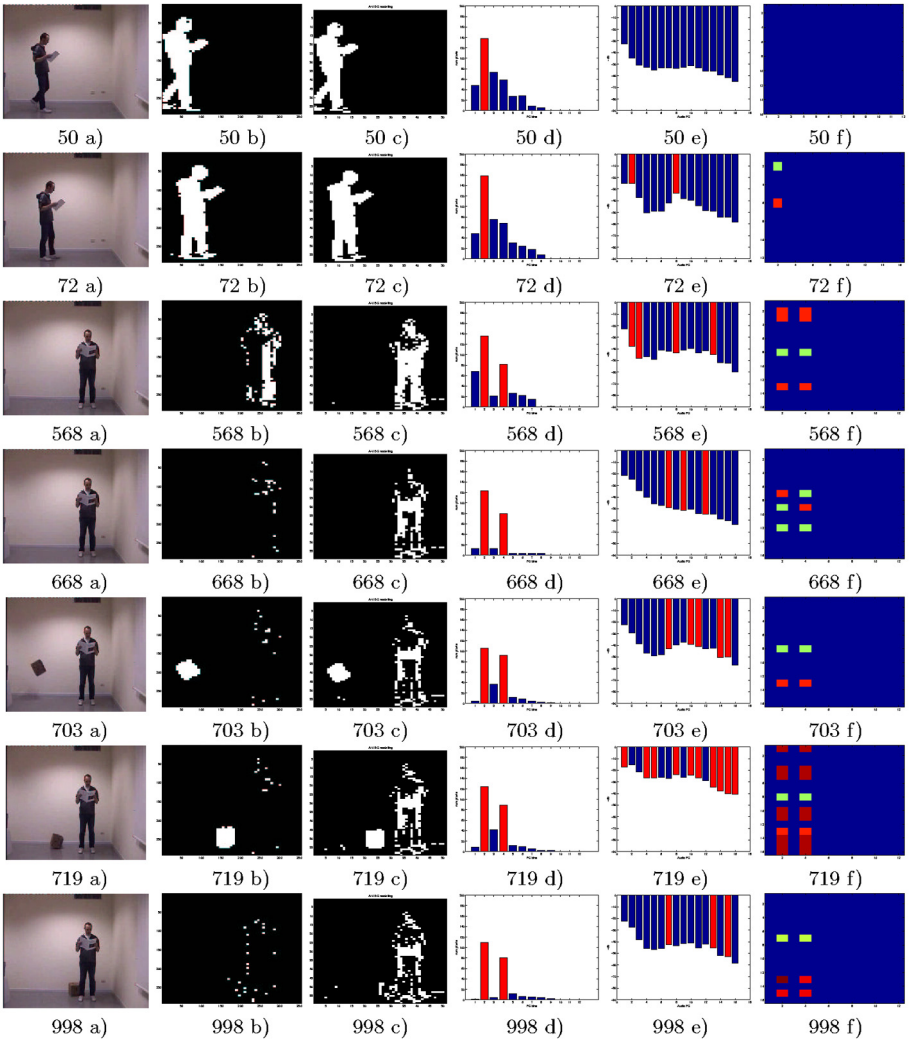

Fig. 2. Comparative results: a) Original sequence; b) Ordinary per pixel video BG modelling; c) Our approach; d) Video novelty detection; e) Audio background modelling; f) Causality matrix at time t;

level of the FG histogram that correctly detects this object as new (Fig.2-50 d) (the lighter bins indicate FG). At frame 72, the person begins to speak, causing an increment of some subbands of the audio spectrum, which is detected as FG by the audio module (Fig. 2-72 e)). Due to the (loose) synchrony of the audio and visual events, the causality matrix evidences a concurrency, as depicted in Fig. 2-72 e). Here, the lightest colored value indicates $\max_i \gamma^{(t)}(i,j)$, i.e., the

maximum causality relation for all audio subbands i, given the video histogram bin j. Therefore, proportionally to the temporal stability of the audio-video FG values, the causality matrix increments some of its entries. Consequently, the learning coefficients of the corresponding audio, histogram, and pixels FG models, become close to zero according to eq. 12. In this way, the synchronized audio and visual FG which remain jointly similar along time are considered as multimodal FG. In the typical video BG modelling scheme, if the visual FG remains still in the scene for a lot of iterations (Fig. 2-568 a) and 668 a)), it loses all its meaning of novelty, so becoming assimilated in the background (Fig. 2- 568 b) and 668 b)). More correctly, in the multimodal case, the FG loses its meaning of novelty only if it remains still without producing sound. In Fig. 2-568 c) and 668 c), the visual aspect of the FG is maintained from the audio FG signal, by exploiting the causality matrix.

The audio visual fusion is also able to preserve the adaptiveness of the BG modelling, if the case. In Fig.2-703 a) and 719 a), a box falls near the talking person, providing new audio and video FG, but, after a while, the box becomes still and silent. In this case, it is correct that it becomes BG after some time (see Fig. 2- 998 b). Also in our approach, the box becomes BG, as the audio pattern decreases quickly, so that no audio visual coupling occurs, and after some iterations the box vanishes, whereas the talking person remains detected (Fig.2- 719 c) and 998 c)). A subtle drawback is notable in Fig.2- 998 c): some parts of box do not completely disappears, because their gray level is similar to that of the talking person, modelled as FG. But this problem could be faced by using a different approach to model visual data (instead of the histogram), or, for instance, a finer quantization of the video histogram space.

5 Conclusions

In this paper, a new concept of multimodal background modelling has been introduced, aimed at integrating audio and video cues for a more robust and complete scene analysis. The separate audio and video streams are modelled using a set of adaptive Gaussian models, able to discover audio and video foregrounds. The integration of audio and video data is obtained posing particular attention to the concept of synchrony, represented using another set of adaptive Gaussian models. The system is able to discover concurrent audio and video cues, which are bound together to define audio-visual patterns. The integrated probabilistic system is able to work on-line using only one camera and one microphone. Preliminary experimental results have shown that this integration permits to face some problems of still video surveillance systems, like the FG sleeping problem.

Acknowledgment. This work was partially supported by the European Commission under the project no. GRD1-2000-25409 named ARROV.

References

1. PAMI: Special issue on video surveillance. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000)
2. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Int. Conf. Computer Vision and Pattern Recognition. Volume 2. (1999)
3. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. In: Int. Conf. Computer Vision. (1999) 255–261
4. Niebur, E., Hsiao, S., Johnson, K.: Synchrony: a neuronal mechanism for attentional selection? Current Opinion in Neurobiology (2002) 190–194
5. Stein, B., Meredith, M.: The Merging of the Senses. MIT Press, Cambridge (1993)
6. Checka, N., Wilson, K.: Person tracking using audio-video sensor fusion. Technical report, MIT Artificial Intelligence Laboratory (2002)
7. Zotkin, D., Duraiswami, R., Davis, L.: Joint audio-visual tracking using particle filters. EURASIP Journal of Applied Signal Processing **2002** (2002) 1154–1164
8. Wilson, K., Checka, N., Demirdjian, D., Darrell, T.: Audio-video array source separation for perceptual user interfaces. In: Proceedings of Workshop on Perceptive User Interfaces. (2001)
9. Darrell, T., Fisher, J., Wilson, K.: Geometric and statistical approaches to audio-visual segmentation for unthetered interaction. Technical report, CLASS Project (2002)
10. Hershey, J., Movellan, J.R.: Audio-vision: Using audio-visual synchrony to locate sounds. In: Advances in Neural Information Processing Systems 12, MIT Press (2000) 813–819
11. Mason, M., Duric, Z.: Using histograms to detect and track objects in color video. In: The 30th IEEE Applied Imagery Pattern Recognition Workshop (AIPR'01), Washington, D.C., USA (2001) 154–159
12. Bregman, A.: Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, London (1990)
13. Peltonen, V.: Computational auditory scene recognition. Master's thesis, Tampere University of Tech., Finland (2001)
14. Cowling, M., R.Sitte: Comparison of techniques for environmental sound recognition. Pattern Recognition Letters (2003) 2895–2907
15. Zhang, T., Kuo, C.: Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing **9** (2001) 441–457
16. Roweis, S.: One microphone source separation. In: Advances in Neural Information Processing Systems. (2000) 793–799
17. Hild II, K., Erdogmus, D., Principe, J.: On-line minimum mutual information method for time-varying blind source separation. In: Intl. Workshop on Independent Component Analysis and Signal Separation (ICA '01). (2001) 126–131
18. Marple, S.: Digital Spectral Analysis. second edn. Prentice-Hall (1987)